09/331-631 #4 Cont

PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 5 : C12N 15/29, C07K 13/00 C12N 1/21, 1/19 — A1

(11) International Publication Number: WO 91/19801

(43) International Publication Date: 26 December 1991 (26.12.91)

(21) International Application Number: PCT/GB91/00914

(22) International Filing Date: 7 June 1991 (07.06.91)

(30) Priority data:
90130016.2 11 June 1990 (11.06.90) GB

(71) Applicant *(for all designated States except US)*: MARS UK LIMITED [GB/GB]; 3D Dundee Road, Slough, Berkshire SL1 4JS (GB).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)* : SPENCER, Margaret, Elizabeth [GB/GB]; 24 Totley Brook Road, Sheffield S17 3QS (GB). HODGE, Rachel [GB/GB]; 16 Mount Avenue, Leicester LE5 3RN (GB). DEAKIN, Edward, Alfred [GB/GB]; The Oakes, Oakes Park, Norton, Sheffield S8 8BA (GB). ASHTON, Sean [GB/GB]; 5 Chamberlain Court, Burncross, Chapeltown, Sheffield S30 4ZU (GB).

(74) Agents: SHEARD, Andrew, Gregory et al.; Kilburn & Strode, 30 John Street, London WC1N 2DD (GB).

(81) Designated States: AT (European patent), AU, BE (European patent), BR, CA, CH (European patent), DE (European patent), DK (European patent), ES (European patent), FI, FR (European patent), GB, GB (European patent), GR (European patent), HU, IT (European patent), JP, KR, LU (European patent), NL (European patent), NO, PL, RO, SE (European patent), SU, US.

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: RECOMBINANT 47 AND 31kD COCOA PROTEINS AND PRECURSOR

(57) Abstract

47 kD and 31 kD proteins, and their 67 kD expression precursor, believed to be the source of peptide flavour precursors in cocoa (*Theobroma cacao*) have been identified. Genes coding for them have been probed, identified and sequenced, and recombinant proteins have been synthesised.

RECOMBINANT 47 AND 31kD COCOA PROTEINS AND PRECURSOR

This invention relates to proteins and nucleic acids derived from or otherwise related to cocoa.

The beans of the cocoa plant (*Theobroma cacao*) are the raw material for cocoa, chocolate and natural cocoa and chocolate flavouring. As described by Rohan ("Processing of Raw Cocoa for the Market", FAO/UN (1963)), raw cocoa beans are extracted from the harvested cocoa pod, from which the placenta is normally removed, the beans are then "fermented" for a period of days, during which the beans are killed and a purple pigment is released from the cotyledons. During fermentation "unknown" compounds are formed which on roasting give rise to characteristic cocoa flavour. Rohan suggests that polyphenols and theobromine are implicated in the flavour precursor formation. After fermentation, the beans are dried, during which time the characteristic brown pigment forms, and they are then stored and shipped.

Biehl *et al*, 1982 investigated proteolysis during anaerobic cocoa seed incubation and identified 26kD and 44kD proteins which accumulated during seed ripening and degraded during germination. Biehl asserted that there were storage proteins and suggested that they may give rise to flavour-specific peptides.

Fritz *et al*, 1985 identified polypeptides of 20kD and 28kD appearing in the cytoplasmic fraction of cocoa seed extracts at about 100 days after pollination. It appears that the 20kD protein is thought to have glyceryl acyltransferase activity.

In spite of the uncertainties in the art, as summarised above, proteins apparently responsible for flavour production in cocoa beans have now been identified. Further, it has been discovered that, in spite of Fritz's caution that "cocoa seed

Theobroma cacao has two primary subspecies, *Th. cacao cacao* and *Th. cacao sphaerocarpum*. While proteins in accordance with the invention may be derived from these subspecies. the invention is not limited solely to these subspecies. For example, many cocoa varieties are hybrids between different species; an example of such a hybrid is the trinitario variety.

The invention also relates to nucleic acid, particularly DNA, coding for the proteins referred to above (whether the primary translation products, the processed proteins or fragments). The invention therefore also provides, in further aspects:

- nucleic acid coding for a 67kD protein of *Th. cacao*, or for a fragment thereof;

- nucleic acid coding for a 47kD protein of *Th. cacao*, or for a fragment thereof;

- nucleic acid coding for a 31kD protein of *Th. cacao*, or for a fragment thereof;

Included in the invention is nucleic acid which is degenerate for the wild type protein and which codes for conservative or other non-deleterious mutants. Nucleic acid which hybridises to the wild type material is also included.

Nucleic acid within the scope of the invention will generally be recombinant nucleic acid and may be in isolated form. Frequently, nucleic acid in accordance with the invention will be incorporated into a vector (whether an expression vector or otherwise) such as a plasmid. Suitable expression vectors will contain an appropriate promoter, depending on the intended expression host. For yeast, an appropriate promoter is the yeast pyruvate kinase (PK) promoter: for bacteria an appropriate promoter is a strong lambda promoter.

Figure 4 shows the relationship between the 67kD protein and seed storage proteins from other plants;

Figure 5 shows a map of plasmid pJLA502;

Figure 6 shows schematically the formation of plasmid pMS900;

Figure 7 shows two yeast expression vectors useful in the present invention; vector A is designed for internal expression and vector B is designed for secreted expression;

Figure 8a shows, in relation to vector A, part of the yeast pyruvate kinase gene showing the vector A cloning site, and the use of *Hin-Nco* linkers to splice in the heterologous gene;

Figure 8b shows, in relation to vector B, part of the yeast alpha-factor signal sequence showing the vector B cloning site, and the use of *Hin-Nco* linkers to create an in-phase fusion;

Figure 9a shows how plasmid pMS900 can be manipulated to produce plasmids pMS901, pMS903, pMS907, pMS908, pMS911, pMS912 and pMS914;

Figure 9b shows how plasmid pMS903 can be manipulated to produce plasmids pMS904, pMS905, pMS906, pMS909 and pMS916;

Figure 10 shows maps of plasmids pMS908, pMS914, pMS912, pMS906, pMS916 and pMS910;

Figure 11 shows the construction of a plasmid to express the 67kD protein by means of the AOX promoter on an integrated vector in *Hansenula polymorpha*; and

Characteristics of the Storage Polypeptides

The solubility characteristics of the 47 kD and 31 kD polypeptides were roughly defined by one or two quick experiments. Dialysis of the polypeptide solution against SDS-free extraction buffer rendered the 47 kD and 31 kD polypeptides insoluble, as judged by their ability to pass through a 0.22 micron membrane. Fast Protein Liquid Chromatography (FPLC) analysis also showed that the 47 kD and 31 kD polypeptides were highly associated after extraction with McIlvaines buffer pH 6.8 (0.2 M disodium hydrogen phosphate titrated with 0.1 M citric acid). The 47 kD and 31 kD polypeptides are globulins on the basis on their solubility.

Purification of the 47 kD and 31 kD polypeptides

The 47 kD and 31 kD polypeptides were purified by two rounds of gel filtration on a SUPEROSE-12 column of the PHARMACIA Fast Protein Liquid Chromatography system (FPLC), or by electroelution of bands after preparative electrophoresis. (The words SUPEROSE and PHARMACIA are trade marks.) Concentrated protein extracts were made from 50 mg acetone powder per ml of extraction buffer, and 1-2 ml loaded onto 2 mm thick SDS-PAGE gels poured without a comb. After electrophoresis the gel was surface stained in aqueous Coomassie Blue, and the 47 kD and 31 kD bands cut out with a scalpel. Gel slices were electroeluted into dialysis bags in electrophoresis running buffer at 15 V for 24 hours, and the dialysate dialysed further against 0.1% SDS. Samples could be concentrated by lyophilisation.

Example 2

Amino-acid Sequence Data from Proteins

Protein samples (about 10 μg) were subjected to conventional N-terminal amino-acid sequencing. The 47 kD and 31 kD polypeptides were N-terminally blocked, so cyanogen bromide peptides of the 47 kD and 31 kD peptides were

The gamma-globulin fraction of the serum was partially purified by precipitation with 50% ammonium sulphate, solubilisation in phosphate-buffered saline (PBS) and chromatography on a DE 52 cellulose ion-exchange column as described by Hill, 1984. Fractions containing gamma-globulin were monitored at 280 nm (OD_{280} of 1.4 is equivalent to 1 mg/ml gamma-globulin) and stored at -20°C.

The effective titre of the antibodies was measured using an enzyme-linked immunosorbant assay (ELISA). The wells of a polystyrene microtitre plate were coated with antigen (10-1000 ng) overnight at 4°C in carbonate coating buffer. Wells were washed in PBS-Tween and the test gamma globulin added at concentrations of 10, 1 and 0.1 μg/ml (approximately 1:100, 1:1000 and 1:10,000 dilutions). The diluent was PBS-Tween containing 2% polyvinyl pyrrolidone (PVP) and 0.2% BSA. Controls were preimmune serum from the same animal. Binding took place at 37°C for 3-4 hours. The wells were washed as above and secondary antibody (goat anti-rabbit IgG conjugated to alkaline phosphatase) added at a concentration of 1 μg/ml, using the same conditions as the primary antibody. The wells are again washed, and alkaline phosphatase substrate (p-nitrophenyl phosphate; 0.6 mg/ml in diethanol-amine buffer pH 9.8) added. The yellow colour, indicating a positive reaction, was allowed to develop for 30 minutes and the reaction stopped with 3M NaOH. The colour is quantified at 405 nm. More detail of this method is given in Hill, 1984. The method confirmed that the antibodies all had a high titre and could be used at 1 μg/ml concentration.

Example 4

Isolation of Total RNA from Immature Cocoa Beans

The starting material for RNA which should contain a high proportion of mRNA specific for the storage proteins was immature cocoa beans, at about 130 days after pollination. Previous work had suggested that synthesis of storage proteins was approaching its height by this date (Biehl *et al.* 1982). The beans are roughly corrugated and pale pinkish-purple at this age.

Preparation of mRNA From Total RNA

The mRNA fraction was separated from total RNA by affinity chromatography on a small (1 ml) oligo-dT column, the mRNA binding to the column by its poly A tail. The RNA (1-2 mg) was denatured by heating at 65°C and applied to the column in a high salt buffer. Poly A+ was eluted with low salt buffer, and collected by ethanol precipitation. The method is essentially that of Aviv and Leder (1972), modified by Maniatis *et al* (1982). From 1 mg of total RNA, approximately 10-20 μg polyA+ RNA was obtained (1-2%).

In vitro *Translation of mRNA*

The ability of mRNA to support *in vitro* translation is a good indication of its cleanliness and intactness. Only mRNAs with an intact polyA tail (3' end) will be selected by the oligo-dT column, and only mRNAs which also have an intact 5' end (translational start) will translate efficiently. *In vitro* translation was carried out using RNA-depleted wheat-germ lysate (Amersham International), the *de novo* protein synthesis being monitored by the incorporation of [^{35}S]-methionine (Roberts and Paterson, 1973). Initially the rate of *de novo* synthesis was measured by the incorporation of [^{35}S]-methionine into TCA-precipitable material trapped on glass fibre filters (GFC, Whatman). The actual products of translation were investigated by running on SDS-PAGE, soaking the gel in fluor, drying the gel and autoradiography. The mRNA preparations translated efficiently and the products covered a wide range of molecular weights, showing that intact mRNAs for even the largest proteins had been obtained. None of the major translation products corresponded in size to the 47kD or 31kD storage polypeptides identified in mature beans, and it was apparent that considerable processing of the nascent polypeptides must occur to give the mature forms.

Example 6

cDNA Synthesis From the mRNA Preparations

cDNA synthesis was carried out using a kit from Amersham International. The first strand of the cDNA is synthesised by the enzyme reverse transcriptase, using the four nucleotide bases found in DNA (dATP, dTTP, dGTP, dCTP) and an oligo-dT primer. The second strand synthesis was by the method of Gubler and Hoffman (1983), whereby the RNA strand is nicked in many positions by RNase H, and the remaining fragments used to prime the replacement synthesis of a new DNA strand directed by the enzyme *E. coli* DNA polymerase I. Any 3' overhanging ends of DNA are filled in using the enzyme T4 polymerase. The whole process was monitored by adding a small proportion of [^{32}P]-dCTP into the initial nucleotide mixture, and measuring the percentage incorporation of label into DNA. Assuming that cold nucleotides are incorporated at the same rate, and that the four bases are incorporated equally, an estimate of the synthesis of cDNA can be obtained. From 1 μg of mRNA approximately 140 ng of cDNA was synthesised. The products were analysed on an alkaline 1.4% agarose gel as described in the Amersham methods. Globin cDNA, synthesised as a control with the kit, was run on the same gel, which was dried down and autoradiographed. The cocoa cDNA had a range of molecular weights, with a substantial amount larger than the 600 bp of the globin cDNA.

Example 7

Cloning of cDNA into a Plasmid Vector by Homopolymer Tailing

The method of cloning cDNA into a plasmid vector was to 3' tail the cDNA with dC residues using the enzyme terminal transferase (Boehringer Corporation Ltd), and anneal into a *Pst*I-cut and 5' tailed plasmid (Maniatis *et al*, 1982 Eschenfeldt *et al*, 1987). The optimum length for the dC tail is 12-20 residues. The tailing reaction (conditions as described by the manufacturers) was tested

```
                Met-Phe-Glu-Ala-Asn-Pro
5'    ATG TTT GAA GCT AAT CC   3'
         C   G   C   C
                 A
                 G
```

The actual probe was made anti-sense so that it could also be used to probe mRNA. Probe synthesis was carried out using an Applied Biosystems apparatus.

Example 9

Use of Oligonucleotides to Probe cDNA Library

The oligonucleotide probes were 5' end-labelled with gamma-[^{32}P] dATP and the enzyme polynucleotide kinase (Amersham International). The method was essentially that of Woods (1982, 1984), except that a smaller amount of isotope (15 μCi) was used to label about 40 ng probe, in 10 mM $MgCl_2$, 100 mM Tris-HCl, pH 7.6; 20 mM 2-mercaptoethanol.

The cDNA library was grown on GeneScreen (New England Nuclear) nylon membranes placed on the surface of L-agar + 100 μg/ml ampicillin plates. (The word GeneScreen is a trade mark.) Colonies were transferred from microtitre plates to the membranes using a 6 x 8 multi-pronged device, designed to fit into the wells of half the microtitre plate. Colonies were grown overnight at 37°C, lysed in sodium hydroxide and bound to membranes as described by Woods (1982, 1984). After drying the membranes were washed extensively in 3 x SSC/0.1% SDS at 65°C, and hybridised to the labelled probe, using a HYBAID apparatus from Hybaid Ltd, PO Box 82, Twickenham, Middlesex. (The word HYBAID is a trade mark.) Conditions for hybridisation were as described by Mason & Williams (1985), a T_d being calculated for each oligonucleotide according to the formula:

replication, and the single-strands are packaged as phages extruded into the medium. DNA can be prepared from these 'phages' using established methods for M13 phages (Miller, 1987), and used for sequencing by the method of Sanger (1977) using the reverse sequencing primer. The superinfecting phage used is a derivative of M13 termed M13K07, which replicates poorly and so does not compete well with the plasmid, and contains a selectable kanamycin-resistance marker. Detailed methods for preparing single-strands from the pTZ plasmids and helper phages are supplied by Pharmacia. DNA sequence was compiled and analysed using the Staden package of programs (Staden, 1986), on a PRIME 9955 computer. (The word PRIME is a trade mark.)

Example 12

Features of the 47 kD/31 kD cDNA and Deduced Amino-acid Sequence of the 67 kD Precursor

DNA sequencing of the three positive clones, pMS600, pMS700, pMS800, confirmed the overlap presumed in Figure 1. No sequence differences were found in the overlapping regions (about 300 bp altogether), suggesting that the three cDNAs were derived from the same gene. The sequence of the combined cDNAs comprising 1818 bases is shown in Figure 2. The first ATG codon is found at position 14, and is followed by an open reading frame of 566 codons. There is a 104-base 3' untranslated region containing a polyadenylation signal at position 1764. The oligonucleotide probe sequence is found at position 569.

The open reading frame translates to give a polypeptide of 566 amino-acids (Figure 2), and a molecular weight of 65612, which is reasonably close to the 67 kD measured on SDS-PAGE gels. The N-terminal residues are clearly hydrophobic and look like a characteristic signal sequence. Applying the rules of Von Heije (1983), which predict cleavage sites for signal sequences, suggests a cleavage point between amino-acids 20 and 21 (see Figure 3). The region following this is highly hydrophobic and contains four Cys-X-X-X-Cys motifs.

Example 13

Expression of the 67 kD Polypeptide in E. coli

Before the 67 kD coding region could be inserted into a expression vector the overlapping fragments from the three separate positive clones had to be spliced into a continuous DNA segment. The method of splicing is illustrated in Figure 6: a *Hin*dIII-*Bgl*II fragment from pMS600, a *Bgl*II-*Eco*RI fragment from pMS700 and an *Eco*RI-*Sal*I fragment from pMS800 were ligated into pTZ19R cut with *Hin*dIII and *Sal*I. The resulting plasmid, containing the entire 67 kD cDNA, was termed pMS900.

An *Nco*I site was introduced at the ATG start codon, using the mutagenic primer:

5' TAG CAA CCA TGG TGA TCA 3'.

In vitro mutagenesis was carried out using a kit marketed by Amersham International, which used the method of Eckstein and co-workers (Taylor *et al*, 1985). After annealing the mutagenic primer to single-stranded DNA the second strand synthesis incorporates alpha-thio-dCTP in place of dCTP. After extension and ligation to form closed circles, the plasmid is digested with *Nci*I, an enzyme which cannot nick DNA containing thio-dC. Thus only the original strand is nicked, and subsequently digested with exonuclease III. The original strand is then resynthesised, primed by the remaining DNA fragments and complementing the mutated position in the original strand. Plasmids are then transformed into *E. coli* and checked by plasmid mini preparations.

The 67 kD cDNA was then cloned into the *E. coli* expression plasmid, pJLA502 (Figure 5), on an *Nco*I - *Sal*I fragment (pMS902).

mating alpha-factor downstream of the promoter, with a *Hind*III site within it to create fusion proteins with incoming coding sequences. The vectors are illustrated in Figure 7.

To use the vectors effectively it is desirable to introduce the foreign coding region such that for vector A, the region from the *Hind*III cloning site to the ATG start is the same as the yeast PK gene, and for vector B, the remainder of the alpha-factor signal, including the lysine at the cleavage point. In practice this situation was achieved by synthesising two sets of *Hind*III - *Nco*I linkers to breach the gap between the *Hind*III cloning site in the vector and the *Nco*I at the ATG start of the coding sequence. This is illustrated in Figure 8.

In order to use the yeast vector B, the hydrophobic signal sequence must first be deleted from the 67 kD cDNA. Although direct evidence of the location of the natural cleavage site was lacking, the algorithm of Von Heije predicts a site between amino-acids 20 (alanine) and 21 (leucine). However it was decided to remove amino-acids 2-19 by deletion, so that the useful *Nco*I site at the translation start would be maintained.

For ease of construction of the yeast vectors, the strategy was to first clone the *Hind*III - *Nco*I linkers into the appropriate pTZ plasmids, and then to clone the linkers plus coding region into the yeast vectors on *Hind*III - *Bam*HI fragments. However the coding region contains an internal *Bam*HI which must be removed by in vitro mutagenesis, giving a new plasmid pMS903. The signal sequence was deleted from pMS903 using the mutagenic primer

5' AGCATAGCAACCATGGTTGCTTTGTTCT 3'

to give pMS904. The appropriate *Hind*III - *Nco*I linkers were then cloned into pMS903 and pMS904 to give pMS907 and pMS905 respectively, and the *Hind*III - *Bam*HI fragments (linkers + coding region) subcloned from these

concentrated 10-25 x in an AMICON mini concentrator. (The word AMICON is a trade mark.) The washed cells were weighed and resuspended in lysis buffer plus protease inhibitors (1 mM phenyl methyl sulphonyl fluoride (PMSF); 1 μg/ml aprotinin; 0.5 μg/ml leupeptin) at a concentration of 1 g/ml. 1 volume acid-washed glass-beads was added and the cells broken by vortexing for 8 minutes in total, in 1 minute bursts, with 1 minute intervals on ice. After checking under the microscope for cell breakage, the mixture was centrifuged at 7000 rpm for 3 minutes to pellet the glass beads. The supernatant was removed to a pre-chilled centrifuge tube, and centrifuged for 1 hour at 20,000 rpm. (Small samples can be centrifuged in a microcentrifuge in the cold.) The supernatant constitutes the soluble fraction. The pellet was resuspended in 1 ml lysis buffer plus 10% SDS and 1% mercaptoethanol and heated at 90°C for 10 minutes. After centrifuging for 15 minutes in a microcentrifuge the supernatant constitutes the particulate fraction.

Samples of each fraction and the concentrated medium were examined by Western blotting. Considering first the plasmids designed for internal expression in YVA, pMS908 produced immunoreactive proteins at 67 kD and 16 kD within the cells only. There was no evidence of the 67 kD protein being secreted under the influence of its own signal sequence. The smaller protein is presumed to be a degradation product. A similar result, but with improved expression, was obtained with pMS914, in which the plant terminator is replaced by a yeast terminator. However in pMS912, in which the coding region for the hydrophilic domain has been deleted, no synthesis of immunoreactive protein occurred.

For industrial production of heterologous proteins in yeast a secreted mode is preferable because yeast cells are very difficult to disrupt, and downstream processing from total cell protein is not easy. The results from the vectors constructed for secreted expressed were rather complicated. From the simplest construct, pMS906, in which the yeast α-factor signal sequence replaces the plant protein's own signal, immunoreactive proteins of approximately 47 kD, 28 kD and 18-20 kD were obtained and secreted into the medium. At first sight

promoter is completely repressed. This means that cells containing the heterologous gene can be grown to a high density on glucose, and induced to produce the foreign protein by allowing the glucose to run out and adding methanol.

A plasmid, pHGL1, containing the MOX promoter and terminator, and a cassette containing the yeast α-factor secretory signal sequence, were prepared. The 67 kD coding region was cloned into pHGL1 on a *Bam*HI - *Bam*HI fragment, replacing the *Bgl*II fragment which contains the 3' end of the MOX coding region. The whole promoter - gene - terminator region can then be transferred to YEp13 on a *Bam*HI - *Bam*HI fragment to give the expression plasmid pMS922. The details of the construction are illustrated in Figure 11. An analogous expression plasmid, pMS925, has been constructed with the yeast α-factor spliced onto the 67 kD coding region, replacing the natural plant signal. The *Bam*HI - *Hind*III cassette containing the α-factor was ligated to the *Hind*III - *Bam*HI fragment used to introduce the 67 kD coding region into YVB. The α-factor plus coding region was then cloned with pHGL1 on a *Bam*HI - *Bam*HI fragment, and transferred into YEP13 as before. Details are shown in Figure 12.

Both constructs have been transformed into *Hansenula* and grown under inducing conditions with 0.5% or 1% methanol. Both constructs directed the production of immunoreactive protein within the cells, and pMS925 secreted the protein into the medium under the influence of the α-factor signal sequence.

E. coli Strains

RR1	$F^-v_B^-M_B$ *ara*-14 *pro*A2 *leu*B6 *lac*Y1 *gal*K2 *vps*L20 (str^r) *xyl*-5 *mtl*-1 *sup*E44
CAG629	lac_{am} tvp_{am} pho_{am} $htpR_{am}$ *mal* *rpsL* *lon* $supC_{ts}$

Cuming, A.C., Williams. R.S., and Cullimore, J.V. in "Immunology in Plant Science", Ed. Wang, T.L., Cambridge University Press, 1986. The use of Antibodies in Molecular Biology.

Doyle, J.J., Schuler, M.A., Godette, W.D., Zencer, V., Beachy, R.N., and Slightom, J.L. *J. Biol. Chem.* 261, 9228-9238 (1986). The glycosylated seed storage proteins of Glycine max and Phaseolus vulgaris : Structural homologies of genes and proteins.

Dretzen, G., Bellard, M., Sassone-Corsi, P., and Chambon, P. *Analytical Biochemistry* 112, 295-298 (1981). A reliable method for the recovery of DNA fragments from agarose and acrylamide gels.

Eschenfeldt, W.H., Puskas, R.S., and Berger, S.L. *Methods in Enzymology* 152, 337-342 (1987). Homopolymeric Tailing.

Fritz *et al* (*J. Food Sci.* 50 946-950 (1985))

Gubler, U., and Hoffman, B.J. *Gene* 25, 263 (1983). A simple and very efficient method for generating cDNA libraries.

Hall, T.C., Ma, Y., Buchbinder, B.U., Pyrne J.W., Sun, S.M., and Bliss, F.A. *Proc. Natl. Acad. Sci. USA* 75, 3196-3200 (1978). Messenger RNA for G1 protein of French bean seeds : cell-free translation and product characterisation.

Hill, S.A. "Methods in Plant Virology", Blackwell 1984.

Johnston, J.R. in "Yeast : A practical approach". Eds Campbell, I., and Duffus, J.H. IRL Press, 1988. Yeast Genetics, Molecular Aspects.

Kreil, G. *Annual Rev. Biochem.* 50, 317-348 (1981). Transfer of proteins across membranes.

Roberts, B.E., and Paterson, B.M. *Proc. Natl. Acad. Sci. USA* **70**, 2330 (1973). Efficient translation of tobacco mosaic virus RNA and rabbit globin 9S RNA in a cell-free system from commercial wheat germ.

Sanger, F., Nicklen, S., and Coulson, A.R. *Proc. Natl. Acad. Sci. USA* **74**, 5463-5467 (1977). DNA sequencing with chain-terminating inhibitors.

Schauder, B., Blocker, H., Frank, R., and McCarthy, J.E.G. *Gene* **52**, 279-283 (1987). Inducible expression vectors incorporating the *E. coli aptE* translation initiation region.

Scott, R., Draper, J., Jefferson, R., Dury, G., and Jacob, L. in "Plant Genetic Transformation and Gene Expression: A Laboratory Manual". Eds. Draper, J., Scott, R., Armitage, P., Walden, R. Blackwell 1988. Analysis of gene organisation and expression in plants.

Staden, R. *Nucleic Acids Res.* **14**, 217-231 (1986). The current status and portability of our sequence handling software.

Sudbery, P.E., Gleeson, M.A., Veale, R.A., Lederboer, A.M., and Zoetmulder, M.C.M. *Biocem. Soc. Trans* 16 1081-103 (1988). Hansenula polymorpha as a novel yeast system for the expression of heterologous genes.

Taylor, J.W., Ott, J., and Eckstein, F. *Nucleic Acids Res.* **13**, 8765-8785 (1985). The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA.

Towbin, H., Staehelin, T., and Gordon, J. *Proc. Natl. Acad. Sci. USA* **76**, 4350-4534 (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets : procedure and some applications.

Von Heije, G. *Eur. J. Biochem* 133, 17-21 (1983). Patterns of Amino-acids near Signal-Sequence Cleavage Sites.

CLAIMS

1. A 67kD protein of *Theobroma cacao*, or a fragment thereof.

2. A 47kD protein of *Th. cacao*, or a fragment thereof.

3. A 31kD protein of *Th. cacao*, or a fragment thereof.

4. A protein as claimed in claim 1, 2 or 3, having at least part of the sequence shown in Figure 2.

5. A fragment as claimed in any one of claims 1 to 4, which comprises at least four amino acids.

6. A protein or fragment as claimed in any one of claims 1 to 6, which is recombinant.

7. Recombinant or isolated nucleic acid coding for a protein or fragment as claimed in any one of claims 1 to 5.

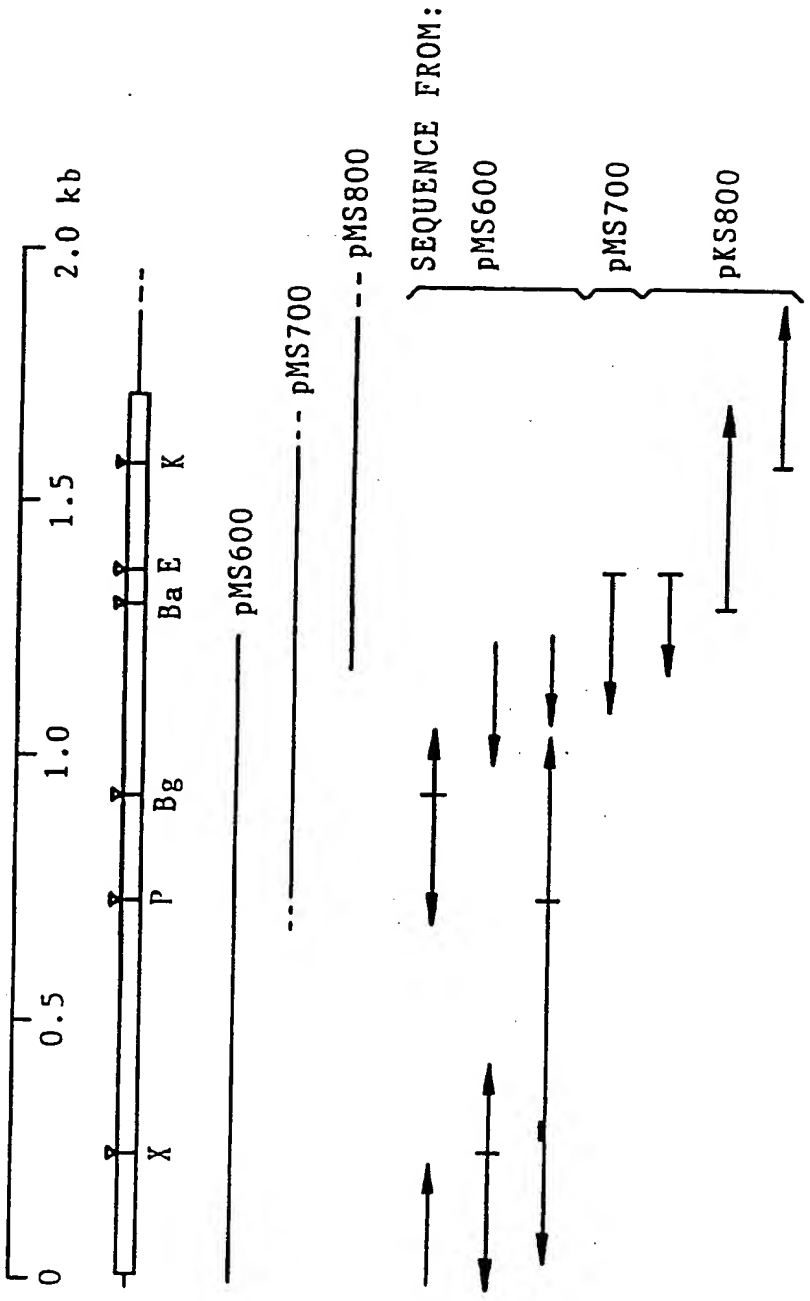
8. Nucleic acid as claimed in claim 7 which is DNA.

9. Nucleic acid as claimed in claim 8, having at least part of the sequence shown in Figure 2.

10. Nucleic acid as claimed in claim 7, 8 or 9, which is in the form of a vector.

11. Nucleic acid as claimed in claim 10, wherein the vector is an expression vector and the protein- or fragment-coding sequence is operably linked to a promoter.

FIG. 1



```
R  D  E  E  G  N  F  K  I  L  Q  R  F  A  E  N  S  P  P  L
CAGGGATGAAGAGGGCAACTTCAAGATCCTCCAGAGGTTTGCTGAGAACTCTCCTCCACT
490       500       510       520       530       540

K  G  I  N  D  Y  R  L  A  M  F  E  A  N  P  N  T  F  I  L
CAAGGGCATCAACGATTACCGCTTGGCCATGTTCGAAGCAAATCCCAACACTTTTATTCT
550       560       570       580       590       600

P  H  H  C  D  A  E  A  I  Y  F  V  T  N  G  K  G  T  I  T
TCCGCACCACTGTGATGCTGAGGCAATTTACTTCGTGACAAACGGAAAGGGACAATTAC
610       620       630       640       650       660

F  V  T  H  E  N  K  E  S  Y  N  V  Q  R  G  T  V  V  S  V
GTTTGTGACTCATGAAAACAAAGAGTCCTATAATGTACAGCGTGGAACAGTAGTCAGCGT
670       680       690       700       710       720

P  A  G  S  T  V  Y  V  V  S  Q  D  N  Q  E  K  L  T  I  A
TCCTGCAGGAAGCACTGTTACGTGGTTAGCCAAGACAACCAAGAGAAGCTAACCATAGC
730       740       750       760       770       780

V  L  A  L  P  V  N  S  P  G  K  Y  E  L  F  F  P  A  G  N
TGTGCTCGCCCTGCCTGTTAATTCTCCTGGCAAATATGAGTTATTCTTCCCCGCTGGAAA
790       800       810       820       830       840

N  K  P  E  S  Y  Y  G  A  F  S  Y  E  V  L  E  T  V  F  N
TAATAAACCTGAATCATATTACGGAGCCTTCAGCTATGAAGTTCTTGAGACCGTCTTCAA
850       860       870       880       890       900

T  Q  R  E  K  L  E  E  I  L  E  E  Q  R  G  Q  K  R  Q  Q
TACACAAGAGAGAAGCTGGAGGAGATCTTGGAGGAACAGAGAGGGCAGAAGAGGCAGCA
910       920       930       940       950       960
```

FIG. 2B

```
  A  S  K  D  Q  P  L  N  A  V  A  F  G  L  N  A  Q  N  N  Q
TGCATCCAAAGACCAGCCCCTGAATGCAGTTGCGTTTGGACTCAACGCCCAGAACAACCA
        1450      1460      1470      1480      1490      1500

  R  I  F  L  A  G  K  K  N  L  V  R  Q  M  D  S  E  A  K  E
GAGAATTTTCCTTGCAGGGAAAAAGAACTTGGTCAGACAAATGGATAGCGAGGCAAAGGA
        1510      1520      1530      1540      1550      1560

  L  S  F  G  V  P  S  K  L  V  D  N  I  F  N  N  P  D  E  S
GTTATCATTTGGGGTACCATCGAAATTGGTAGATAATATATTCAACAACCCGGATGAGTC
        1570      1580      1590      1600      1610      1620

  Y  F  M  S  F  S  Q  Q  R  Q  R  R  D  E  R  R  G  N  P  L
GTATTTCATGTCTTTCTCTCAACAGAGGCAGCGTCGAGATGAAAGGAGGGGCAATCCCTT
        1630      1640      1650      1660      1670      1680

  A  S  I  L  D  F  A  R  L  F  *
GGCCTCAATTCTGGACTTTGCCCGCTTGTTCTAAGCAGCTGCTTCCACTTTTGTATCAGA
        1690      1700      1710      1720      1730      1740

CATGCAGAGGCATGTAATGCAATAAATAAGTTGGCCTATGTAAAGAGGAGAGAGTTTGCT
        1750      1760      1770      1780      1790      1800

TTTGTCTTGTTCTAACCTTGTTTTGAACTAGTAAACTTTCAATGTAATGAGAGTTGTTAT
        1810      1820      1830      1840      1850      1860

CTTTCTA
```

FIG. 2D

FIG. 4A

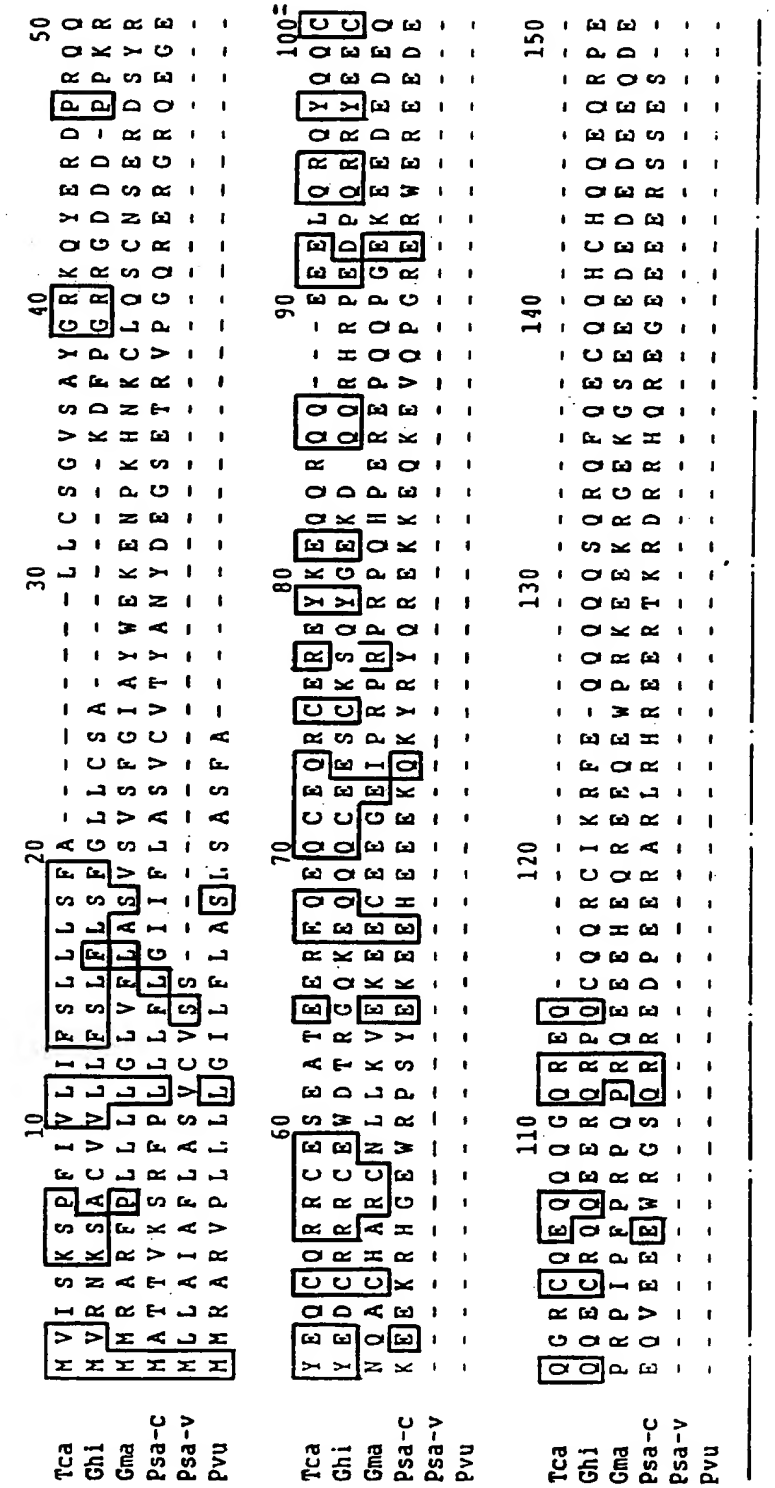
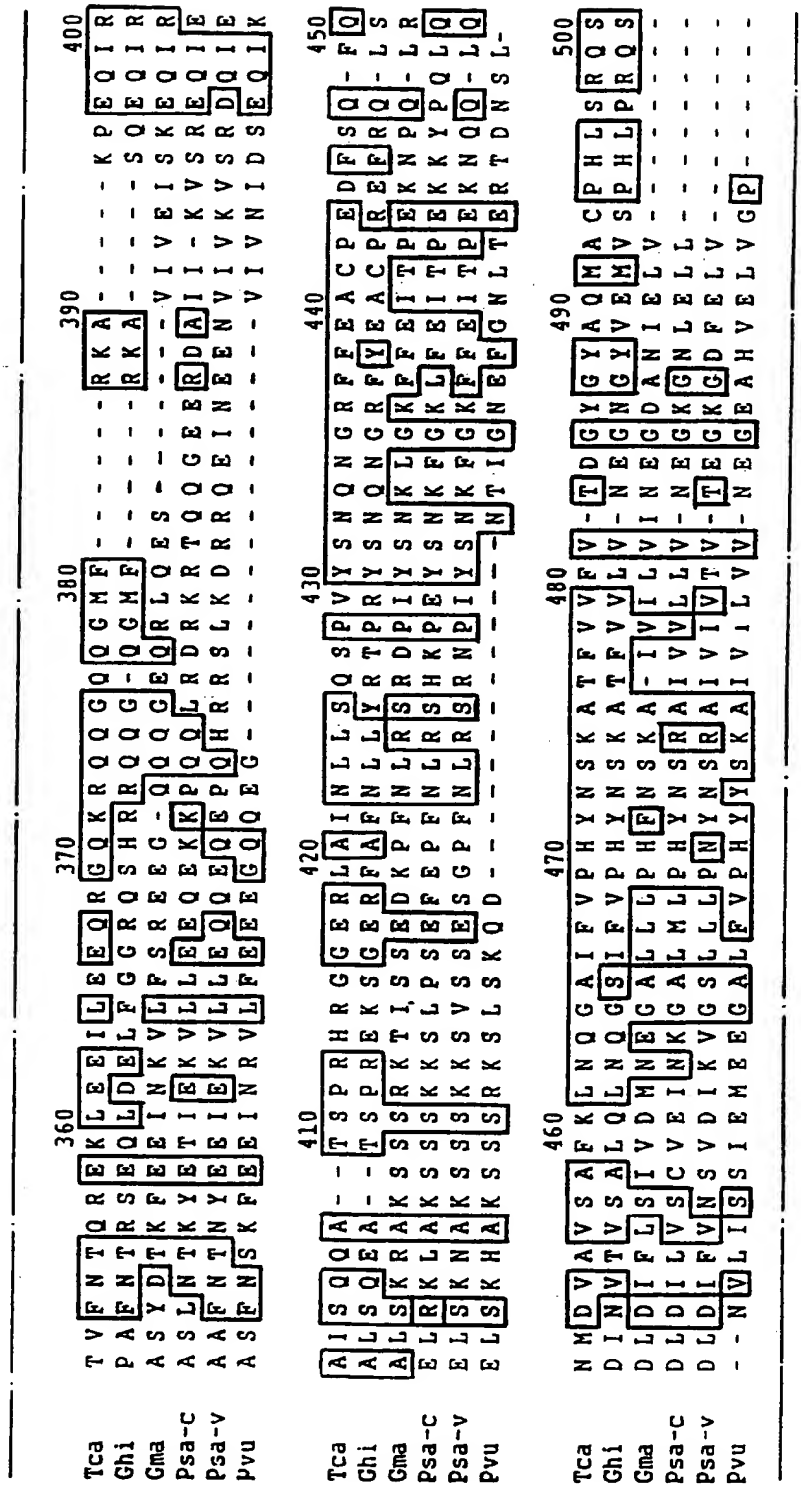
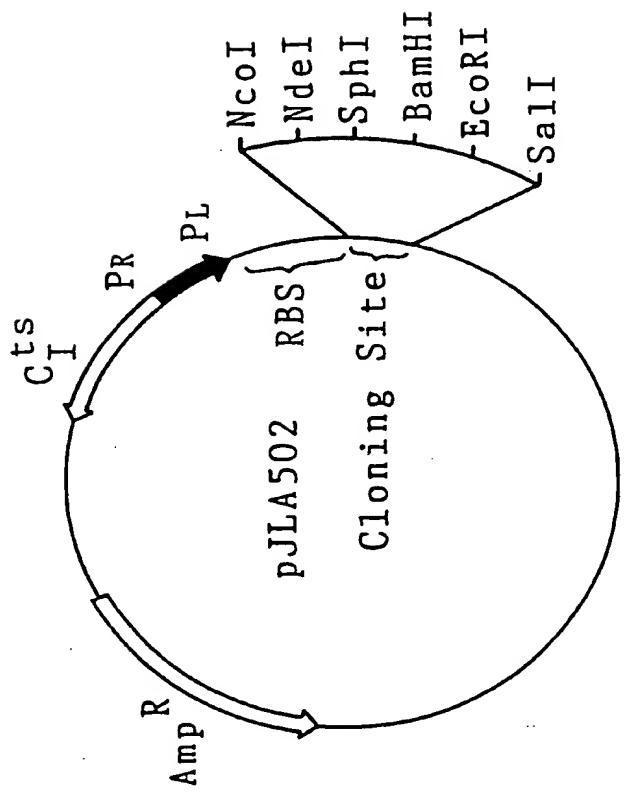


FIG. 4C





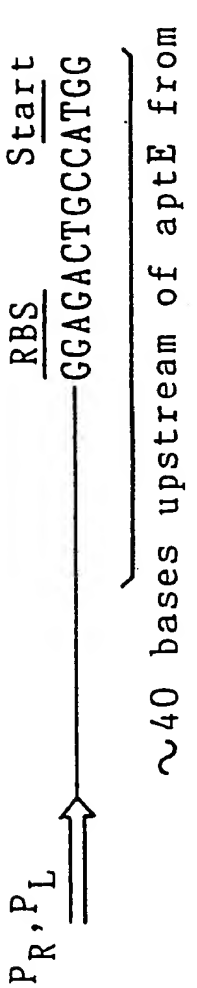


FIG. 5

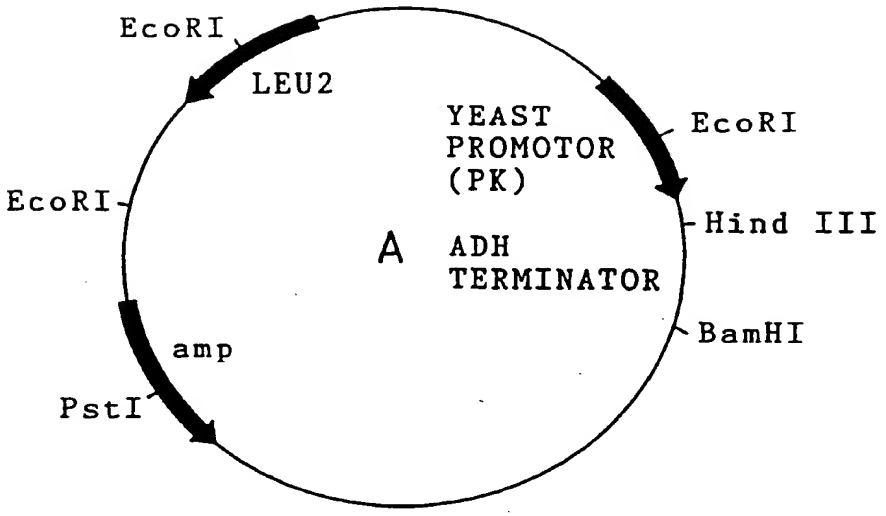


FIG. 7

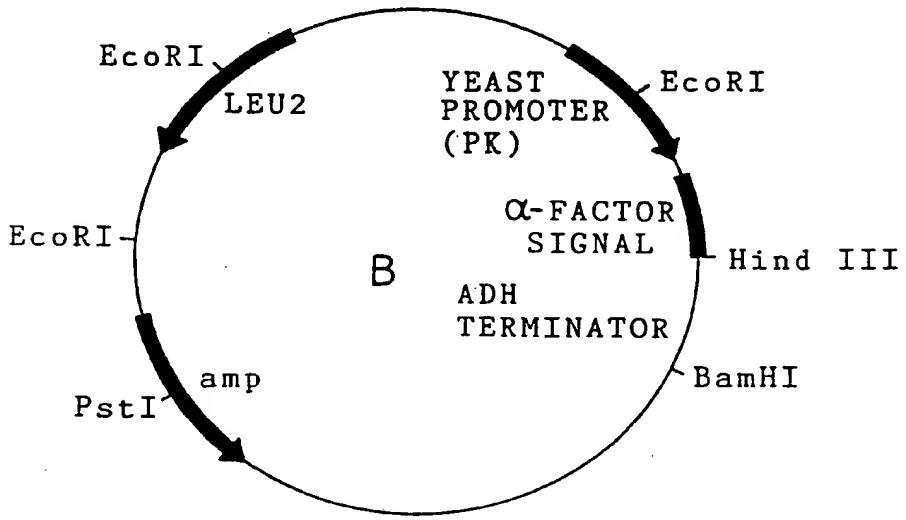


FIG. 8B

YEAST ALPHA-FACTOR SIGNAL SEQUENCE

```
1     231
Met----GluGlyValSerLeuAspLysArgGlu
ATG----GAAGGGGTAAGCTTGGATAAAAGAG
          Hin
```

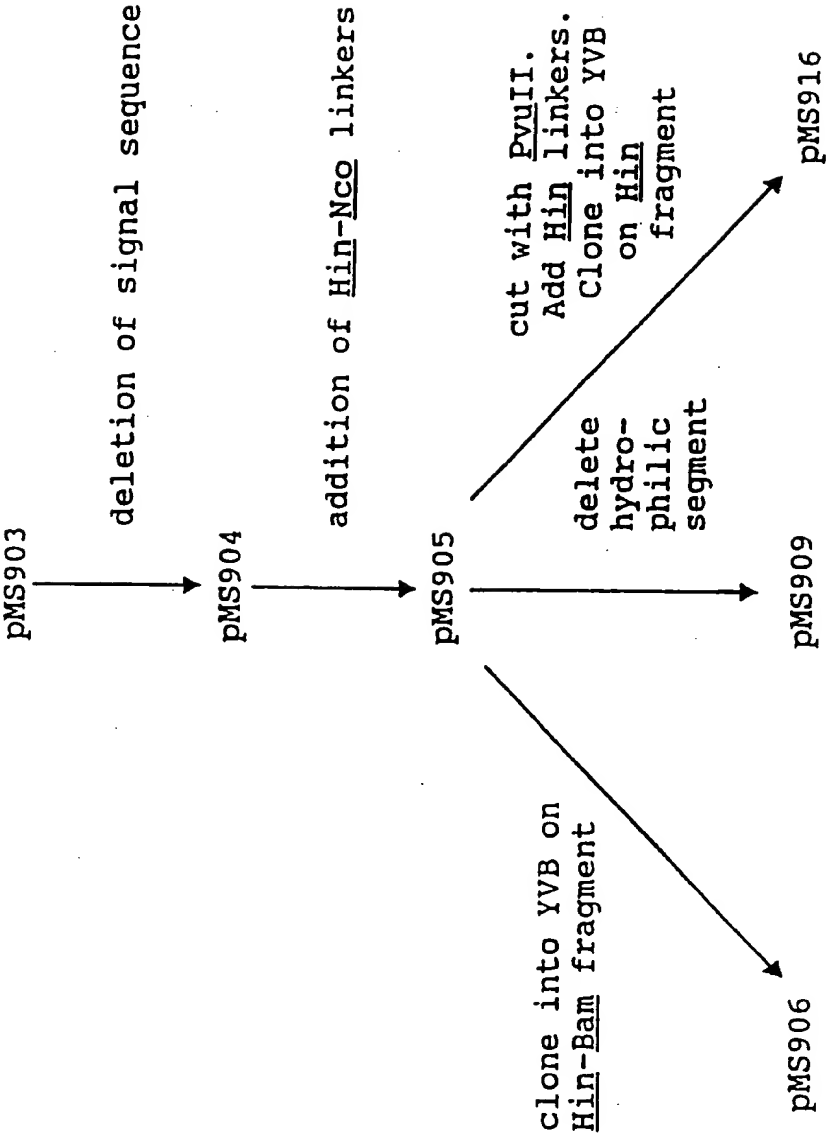
HIN-NCO LINKERS

```
AGCTTGGATAAAAGAGC
ACCTATTTTCTCGGTAC
```

IN-PHASE FUSION OF 67KD CODING REGION

```
Met---GluGlyValSerLeuAspLysArgAlaMetAlaLeu
ATG---GAAGGGGTAAGCTTGGATAAAAGAGCCATGGCGTTG
          Hin          Nco
```

FIG. 9B



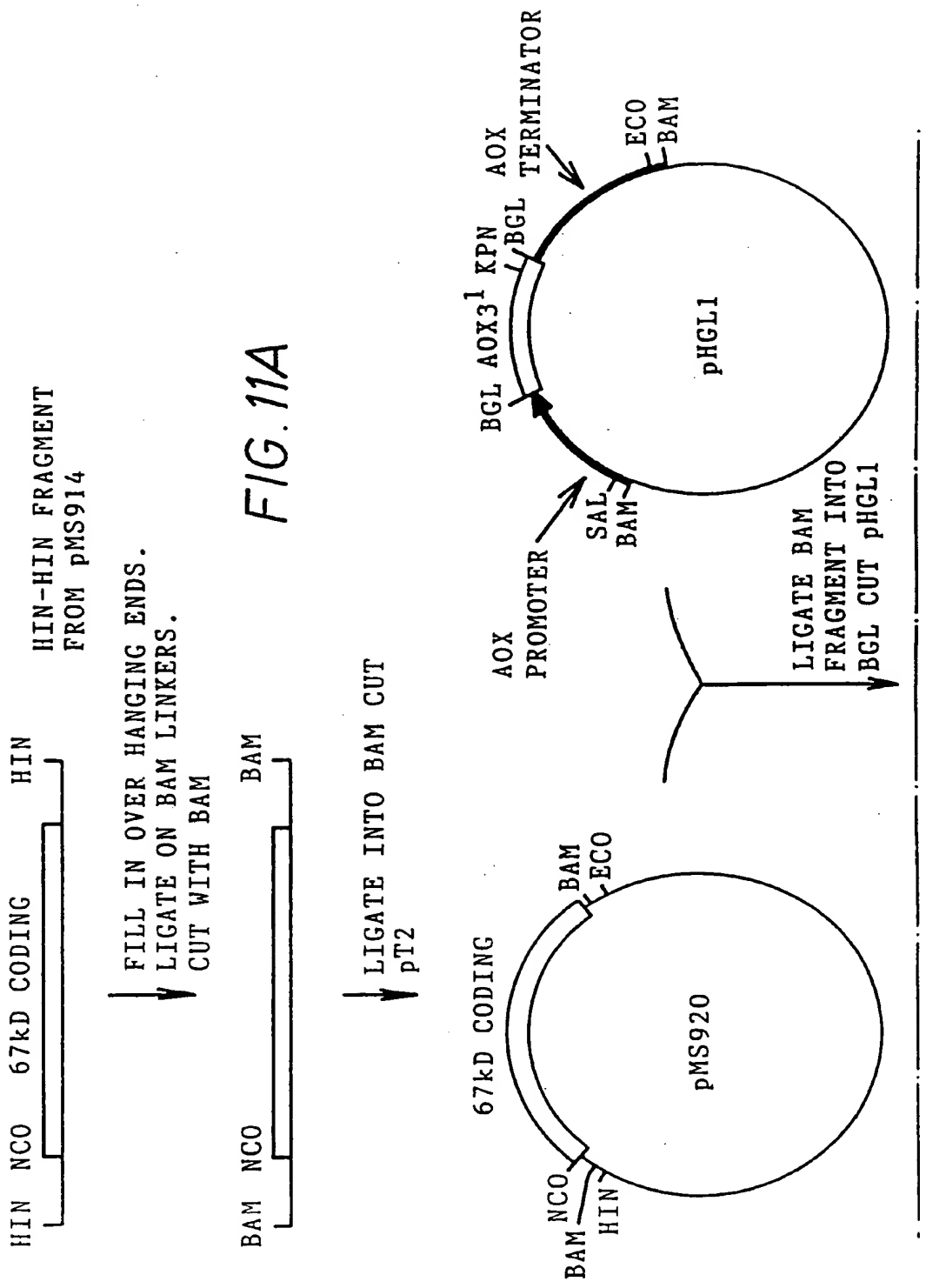


FIG. 11A

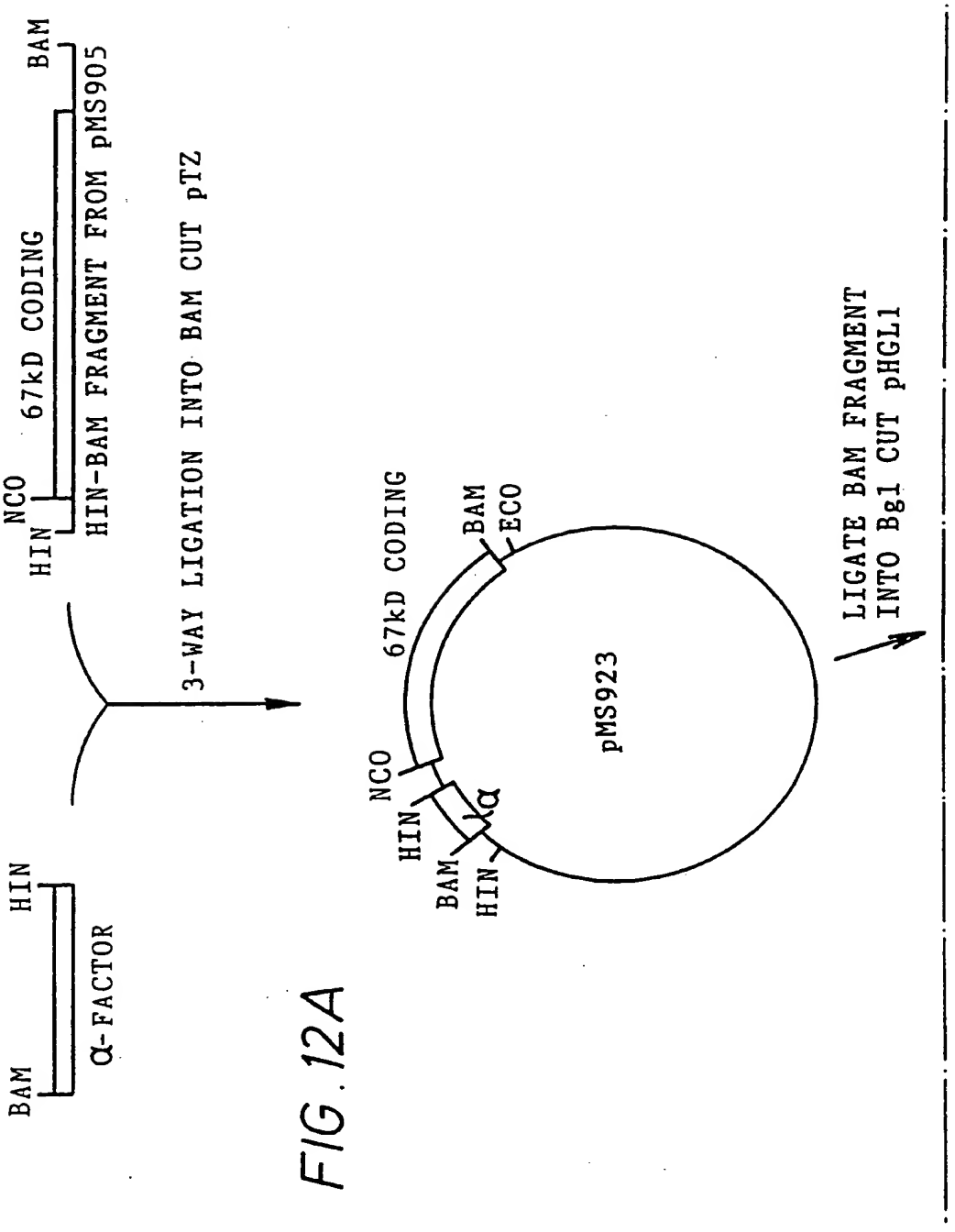


FIG. 12A

INTERNATIONAL SEARCH REPORT

International Application No PCT/GB 91/00914

I. CLASSIFICATION OF SUBJECT MATTER (if several classification symbols apply, indicate all)[6]

According to International Patent Classification (IPC) or to both National Classification and IPC

Int.Cl. 5 C12N15/29 ; C07K13/00 ; C12N1/21 ; C12N1/19

II. FIELDS SEARCHED

Minimum Documentation Searched[7]

Classification System	Classification Symbols
Int.Cl. 5	C07K ; C12N

Documentation Searched other than Minimum Documentation to the Extent that such Documents are Included in the Fields Searched[8]

III. DOCUMENTS CONSIDERED TO BE RELEVANT[9]

Category *	Citation of Document,[11] with indication, where appropriate, of the relevant passages[12]	Relevant to Claim No.[13]
X	J. SCI. FOOD AGRIC. vol. 33, 1982, pages 1291 - 1304; BIEHL, B., ET AL: 'Vacuolar storage proteins of cocoa seeds and their degradation during germination and fermentation ' see the whole document	1-6
X	J. FOOD SCIENCE vol. 50, 1985, pages 946 - 950; FRITZ, P. J., ET AL: 'Cocoa seeds: Changes in protein and polysomal RNA during development ' see the whole document	1-6

-/--

* Special categories of cited documents :[10]

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

IV. CERTIFICATION

Date of the Actual Completion of the International Search	Date of Mailing of this International Search Report
30 SEPTEMBER 1991	20. 11. 91
International Searching Authority	**Signature of Authorized Officer**
EUROPEAN PATENT OFFICE	MADDOX A.D.

Form PCT/ISA/210 (second sheet) (January 1985)

III. DOCUMENTS CONSIDERED TO BE RELEVANT (CONTINUED FROM THE SECOND SHEET)

Category *	Citation of Document, with indication, where appropriate, of the relevant passages	Relevant to Claim No.
X	PLANT MOL BIOL vol. 11, 1988, pages 683 - 695; HIGGINS T.J.V., ET.AL.: 'The sequence of a pea vicilin gene and its expression in tansgenic tobacco plants ' see page 684 methods section ---	4-11,15, 17,19